PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/57, 9/64, C12P 21/00, C07K 16/40, C12N 15/11, C12Q 1/68, G01N 33/50

A2

(11) International Publication Number: WO 98/22597

(43) International Publication Date: 28 May 1998 (28.05.98)

(21) International Application Number: PCT/US97/21684

(22) International Filing Date: 20 November 1997 (20.11.97)

(30) Priority Data:

60/031,196	20 November 1996 (20.11.96)	US
60/046,126	9 May 1997 (09.05.97)	US

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(71) Applicant: OKLAHOMA MEDICAL RESEARCH FOUNDATION [US/US]; 825 N.E. Thirteenth Street, Oklahoma City, OK 73104 (US).

(72) Inventors: KEOLSCH, Gerald; 11617 Roxboro Avenue, Oklahoma City, OK 73162 (US). LIN, Xinli; 1201 Charlton, Edmond, OK 73003 (US). TANG, Jordan, J., N.; 1204 Leawood Drive, Edmond, OK 73034 (US).

(74) Agents: PABST, Patrea, L. et al.; Arnall Golden & Gregory, 2800 One Atlantic Center, 1201 West Peachtree Street, Atlanta, GA 30309-3450 (US).

(54) Title: CLONING AND CHARACTERIZATION OF NAPSIN, AN ASPARTIC PROTEASE

(57) Abstract

A previously unknown aspartic protease capable of cleavage of proteins by hydrolysis, referred to herein as "napsin", has been cloned from a human liver library. Two cDNA clones have been cloned, sequenced and expressed. These encode isozymes of the protease, referred to as "napsin A" and "napsin B". The gene has also be obtained and partially sequenced. A process for rapid purification of the enzyme using immobilized petpstatin has also been developed, and enzyme isolated from human kidney tissue. Polyclonal antibodies to the enzymes have been made which are also useful for isolation and detection of the enzyme. Similarities to other aspartic proteases, especially cathepsin D, establish the usefulness of the enzyme in diagnostic assays as well as a protease. Either or both the amount or type of napsin expressed in a particular tissue can be determined using labelled antibodies or nucleotide probes to the napsin.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

CLONING AND CHARACTERIZATION OF NAPSIN, AN ASPARTIC PROTEASE

Background of the Invention

The present invention relates to a previously unknown aspartic protease present in human liver, isolated by cloning of a gene from a human liver cDNA library.

Members of the aspartic protease family are characterized by the presence of catalytic aspartic acid residues in their active center. There are five aspartic proteases known to be present in human body. Pepsin and gastricsin are secreted into the stomach for food digestion. Gastricsin is also present in the seminal plasma. Cathepsin D and cathepsin E are present intracellularly to carry out protein catabolism. Renin, which is present in the plasma, is the key enzyme regulating the angiotensin system and ultimately the blood pressure.

Eukaryotic, including human, aspartic proteases are homologous in protein and gene sequences, but have different amino acid and nucleotide sequences. The cDNA and genes of all five human aspartic proteases have been cloned and sequenced. They are synthesized as a single chain zymogen of about 380 residues, which are either secreted or directed to intracellular vacuoles. Upon activation by a self-catalyzed process (except prorenin), an N-terminal *pro* segment of about 45-residues is cleaved off to produce mature enzymes (Tang and Wong, J. Cell. Biochem. 33, 53-63 (1987)). In some cases, for example, with cathepsin D and renin, mature proteases are further cut into two chains. The three-dimensional structures of the aspartic proteases are very similar. Each enzyme contains two internally homologous lobes (Tang *et al.*, Nature 271, 618-621 (1978)). The active-site cleft, which can accommodate eight substrate residues, and two catalytic aspartic acids, are located between the lobes.

These proteases have distinct and important physiological roles. In addition to their importance in physiological functions, these enzymes are also associated with pathological states. For example, human pepsin and

gastricsin are diagnostic indicators for stomach ulcer and cancer (Samloff, Gastroenterology 96, 586-595 (1989); Miki *et al.*, Jpn. J. Cancer Res. 84, 1086-1090 (1993)). Cathepsin D is located in the lysosome. Its main function is the catabolism of tissue proteins. Recent evidence from mice without a functional cathepsin D gene, however, indicates that this enzyme plays a role in the development of intestine in newborn animals. Cathepsin D is also associated with human breast cancer metastasis (Rochefort, Acta Oncologica 31, 125-130 (1992)). Cathepsin E is located in the endoplasmic reticulum of some cells, such as erythrocyte and stomach mucosa cells. It has been applied in the processing of antigens in the immune cells.

Human aspartic proteases have important medical uses. The levels of the proenzymes of human pepsinogen and progastricisin present in the bloodstream and the ratio between the two levels is used in the diagnostic screening of human stomach cancer (Defize, et al., Cancer 59, 952-958 (1987); Miki, et al., Jpn. J. Cancer Res. 84, 1086-1090 (1993)) and ulcer (Miki, et al., Adv. Exp. Med. Biol. 362, 139-143 (1995)). The secretion of procathepsin D is elevated in breast cancer tissue. Thus, the level of procathepsin D in breast cancer is used for clinical prognosis (Rochefort, Acta Oncologica 31, 125-130 (1992)). The analysis of renin in the diagnosis of hypertension is a routine clinical procedure (Brown *et al.*, Handbook of Hypertension 1, 278-323 Robertson, editor (Elsevier Science Publishers, Amsterdam, 1983).

These examples establish that human aspartic proteases are related to human diseases and additional, previously unidentified aspartic proteases, are likely to have clinical applications.

It is therefore an object of the present invention to provide a previously unidentified aspartic protease.

It is a further object of the present invention to characterize and to clone the aspartic protease.

It is still another object of the present invention to identify the tissues in which the aspartic protease is expressed and applications in clinical chemistry and diagnostics.

Summary of the Invention

A previously unknown aspartic protease capable of cleavage of proteins by hydrolysis, referred to herein as "napsin", has been cloned from a human liver library. Two cDNA clones have been cloned, sequenced and expressed. These encode isozymes of the protease, referred to as "napsin A" and "napsin B". One clone is unusual in that it does not include a stop codon but can be used to express protein. The gene has also be obtained and partially sequenced. A process for rapid purification of the enzyme using immobilized petpstatin has also been developed, and enzyme isolated from human kidney tissue. Polyclonal antibodies to the enzymes have been made which are also useful for isolation and detection of the enzyme.

Similarities to other aspartic proteases, especially cathepsin D, establish the usefulness of the enzyme in diagnostic assays as well as as a protease. Either or both the amount or type of napsin expressed in a particular tissue can be determined using labelled antibodies or nucleotide probes to the napsin.

Brief Description of the Drawings

Figure 1 is the cDNA (SEQ ID No. 1) and putative amino acid sequence (SEQ ID No. 2) of human Napsin A. Characteristic active site elements (DTG) and Tyr75 are underlined. The RGD integrin binding motif is also underlined. Lysines at the carboxy terminus correspond to the poly-A region.

Figure 2a is a comparison of the human napsin A amino acid sequence with the amino acid sequences of mouse aspartic protease-like protein (Mori, et al., 1997) and human cathepsin D ("cath D"). Figure

2b is a schematic or dendrogram presentation of sequence relatedness between napsin and other human aspartic proteases.

Figure 3a is the genomic DNA (SEQ ID No. 3) of human Napsin A. Introns are indicated in lower-case letter, exons in upper case. Putative amino acid sequence indicates position of intro-exon junctions. Figure 3b is a schematic presentation of the human napsin A. The exons are shown as vertical bars with the numbering above. The double-headed arrows represent the areas where sequence was determined. The letters are positions of restriction sites where X is XhoI, B is BamHI, and E is EcoRI.

Figure 4 is the cDNA (SEQ ID No. 4) and putative amino acid sequence (SEQ ID No. 5) of human Napsin B. Characteristic active site elements (DTG) and Tyr75 are underlined. The RGD integrin binding motif is also underlined. Lysines at the carboxy terminus correspond to the poly-A region.

Detailed Description of the Invention

I. Cloning and Expression of Napsin Isoforms.

A. Human Napsin A.

1. *Cloning of cDNA encoding Napsin A.*

Clones identified by a homology search of the human cDNA sequence database of the Institute for Genome Research (Adams et al., Science 252, 1651-1656 (1991), reported to encode portions of cathepsin D, were obtained from the American Type Culture Collection, Rockville, MD. These are referred to as ATCC clone number 559204, 540096, 346769, 351669, and 314203; Genbank numbers W19120, N45144, R18106, R11458, and T54068, respectively. Analysis of the sequences indicated these did not encode cathepsin D, and were not full length cDNAs. Primers were designed and used with PCR to obtain additional clones, using a human liver cDNA library as the template. The clones that were obtained include regions not present in the ATCC clones.

Since these clones together provided only about 600 bp of the cDNA, a longer cDNA clone was sought using 5' RACE PCR

(polymerase chain reaction), in which DNA from two separate human liver cDNA libraries cloned into λgt10 was used as template and the primers were based on the near 5'-end sequence (AGGGCACACTGAAGAAGTGGCATCTCC) (SEQ ID No. 5) and the sequence of the λgt10 vector upstream from the insert in the forward direction (CTTTTGAGCAAGTTCAGCCTGGTTAAG) (SEQ ID No. 6). Two clones, pHL-1 (154 bp) and pHL-2 (288 bp) were obtained, one (pHL-2) of which extended the 5'-end sequence into the leader peptide region (Figure 1).

Human napsin A cDNA sequence lacks a stop codon from all clones obtained, yet all features otherwise indicate a functional aspartic protease, including intact active site elements, a conserved Tyr75 (pepsin numbering), and a pro-peptide of approximately 40 amino acids. Different from pepsin, the characteristic aspartic protease, napsin A contains a C-terminal extension, abundance of proline residues, and an RGD motif (integrin-binding motif) near the surface of the 3-D structure of napsin as judged by homologous crystal structures of mammalian aspartic proteases (i.e., pepsin and cathepsin D).

Several related cDNA clones of napsin were obtained by screening of a human liver cDNA library and the nucleotide sequences determined. These clones represent different parts of napsin messenger RNA. Spliced together, the nucleotide sequence encoding napsin A (SEQ ID No. 1) having the deduced amino acid sequence (SEQ ID No. 2) is shown in Figure 1.

2. Expression of Recombinant Napsin A

The cDNA of napsin A, including the leader peptide and the 3' untranslated region and a stretch of polyadenine, was PCR amplified with primers PLHNAP-FWD (SEQ ID No. 7) (5'- AAGCTTATGTCTCCACCACCGCTGCTGCTACCCTTGCTGC) and PLHNAP-REV (SEQ ID No. 8) (5'- AAGCTTTTATTTTTTTTTTTTTTTTTCAATGGAAATATTGG)

and cloned into the HindIII site of vector pLNCX for expression from the CMV promoter (Dusty Miller). Isolated plasmid was transformed into human kidney 293 cells (ATCC). Cells were recovered (8 - 120 mg) and lysed with 50 mM NaOAc, 20 mM zwittergent, pH 3.5 (NAZ buffer) with vortexing. Lysate was incubated on ice for 1 hour. The supernatant from centrifugation at 14,000 xg was employed directly for detection of expressed Napsin A by addition of a 40 μl aliquot of pepstatin-A-agarose (Sigma). The sample was rotated in a 50 ml conical tube at 4°C for 1 week. The matrix was settled and washed twice with 20 ml of NAZ buffer, and three times with 20 mM Tris HCl, 0.5 M KCl, pH 8.2 (TK buffer). Final washes were performed with 20 mM Tris HCl, 50 mM NaCl, and 20 mM zwittergent, pH 9.5. The settled pepstatin-A-agarose (approximately 40 μl) was mixed with 40 μl of SDS-β-mercaptoethanol sample buffer (NOVEX) and heated to 70°C for 10 minutes. Aliquots were applied to 10% Tricine SDS-PAGE (NOVEX) and transblotted to PVDF membranes using a Tris-Tricine buffer system. Membranes were either stained with amido black or blocked with 5% skim milk solution for immunochemical detection. Sections of membrane stained with amido black were excised and washed in sterile H_2O for amino-terminal sequence analysis in an automated Protein Sequencer.

3. *Cloning of Genomic DNA.*

Genomic clones of human napsin were obtained by screening of a human genomic DNA library, cloned into bacterial artificial chromosomes (pBELO-BAC11) (Kim et al., Nucl. Acids Res. 20, 1083-1085 (1992)). The source of genomic DNA for the library was from 978SK and human sperm cell lines, and contained over 140,000 clones. Synthetic oligonucleotide probes were labelled with ^{32}P:

for primary screen Nap-3'

(GAGGGCGAGCGCGCGCCAGTCCCACTCGTGCGCCGCTCTTCATG TCCCCG) (SEQ ID No. 8),

and for secondary screening Nap-5' (CCATCCCCTCAGTAGGTTCAGGGTCCTGCGTCCAGGGTGGACTT GACGAA) (SEQ ID No. 9).

The screening was carried out at Research Genetics, Huntsville, Alabama. Two independent clones were isolated, both approximately 30 kbp in length, and were cut with restriction enzyme and analyzed by pulse-field agarose gel electrophoresis. Fragments of interest were identified by Southern blotting, subcloned into pBlue, and sequenced. The genomic DNA of human Napsin A is shown in Figure 3A.

The human napsin A gene is encoded in 9 exons (Figure 3b). The exon/intron junctions are clearly defined by both the cDNA sequence and the junction motifs. The human napsin A coding region contains an open reading frame starting from the initiation codon ATG (nucleotide 1 in Figure 1) for about 1.2 kb to a polyA stretch in the cDNA sequences. As in the cDNA sequence of napsin A, the genomic exon sequence of napsin A do not contain an in-frame stop codon in the entire coding region before the polyA stretch. The absence of a stop codon in napsin A is confirmed. The absence of stop codon has not been observed for the gene of other mammalian proteins. The cDNA (thus the mRNA) of napsin A is present in different human tissues. It was of interest to see if napsin A gene is capable of expressing protein product. These results are described below.

B. Human Napsin B.

1. cDNA and gene structure.

Clones 559204 and 163167 expressing human napsin B were obtained from ATCC and partially sequenced as described above. Figure 4 displays the resulting full-length DNA sequence encoding Napsin B (SEQ ID No. 3) and the predicted amino acid sequence (SEQ ID No. 4). Nucleotides 1 - 1191 were obtained from genomic clones (described above for Napsin A) and from 1192 - 1910 from ATCC cDNA clones. The napsin B gene sequence is 92% identical to that of napsin A, and the putative protein sequence from each exhibits 91% identity. Similar to

napsin A, the deduced napsin B protein sequence possesses typical aspartic protease motifs, and the same c-terminal extension, RGD motif, and proline-rich regions as in the cDNA of napsin A (Fig. 4). Unlike the napsin A gene, napsin B gene has an in-frame stop codon.

II. Isolation and Characterization of Napsin Protein.

The comparison of the napsin A sequence with three other human aspartic protease proenzymes is shown in Figure 2A. It is clear that napsin is related to human cathepsin D, and is similar to mouse aspartic protease-like protein, but the differences are readily apparent. The relationship to other human aspartic proteases is further analyzed in Figure 2B, which is a diagram of degree of relatedness and also presents the percentage of identical residues. Clearly, by both criteria, napsin differs as much from other aspartic proteases as they differ from one another.

In addition to the sequence similarity to the other human aspartic proteases, the conclusion that napsin is an aspartic protease is drawn from the following observations. (a) The critical active site aspartic residues at positions 32 and 215 are present in the conversed DTG sequences. (b) The presence of Tyr-75 (Y) and some conserved residues around it indicate a functional 'flap' which is characteristic of aspartic proteases. (c) The pro region corresponding to residues 1p to 44p is present in napsin, indicating that it is a proenzyme of the aspartic protease and is capable of activation.

An RGD sequence is found at position 315 to 317 (porcine pepsin residue numbers by convention). This motif has been shown to be important in integrin bonding which is related to the regulation of cellular functions such as cell cycle, hemostasis, inflammation and cell proliferation. This sequence may have particular functional meaning for napsin.

2. *Immunochemical Detection of Napsin A.*

A napsin-specific polyclonal antiserum was produced using the following procedure. An 18 amino acid epitope of Napsin A which was

synthesized as a multiple antignic peptide (MAP) on a poly-lysine backbone by the Molecular Biolgy Resource Facility (OUHSC). This epitope (MKSGARVGLARARPRG) was common to both napsin A and B, and sufficiently dissimilar from cathepsin D, their closest homolog. This region is likely to be located on the surface of Napsin A as determined from the cathepsin D crystal structure coordinates (Erickson, 1993). Aliquots of 1 mg in 1 ml of H_2O were used to immunize goats (Hybridoma Lab, Oklahoma Medical Research Foundation). Serum collected was ammonium sulfate precipated multiple times (Antibodies Lab manual) and affinity purified using the Napsin A MAP coupled to affi-gel 10 (BioRad). This anti sera was used at 1:5000 dilution in the detection of Napsin A on PVDF membranes transblotted from SDS-PAGE gels (NOVEX). The ECL system (Pierce) was used for detection of primary antibody.

Immunoblots of recombinant Napsin A sample from human kidney 293 cells prepared as described above detected Napsin A. These results show expression of napsin A gene produced an immunospecific band which migrated in SDS-polyacrylamide electrophoresis with a similar mobility to that of napsin B. Thus, despite of the absence of a stop codon in napsin A, its protein is correctly expressed in a human cell line. The fact that this napsin A protein was recovered from the pepstatin-affinity column suggests that the presence of an active site similar to all aspartic proteases.

3. Detection of Napsin B in Human Tissue and Cell Lines

Sections of approximately 8 grams of human kidney cortex (Cooperative Human Tissue Network, National Cancer Institute, NIH) were homogenized in a Waring blender in buffer composed of 20 mM Tris HCl, 50 mM NaCl, 20 mM zwittergent, and 1 μM each of TPCK, TLCK, and EDTA, pH 7.5 (buffer TZ). The homogenate was made 40% ammonium sulfate with gentle stirring, and centrifuged 10,000 xg. The resulting supernatant was made 70% ammonium sulfate and centrifuged

10,000 xg. The material insoluble in 70% ammonium sulfate (the 40-70% cut) was dissolved in 15 ml of buffer TZ and made pH 4.0 with 30 ml of NAZ buffer. Following incubation on ice for 1 hour, the sample was centrifuged at 14,000 x g. To the resulting supernatant, a 0.1 ml aliquot of pepstatin-A-agarose (Sigma) was added. Detection of napsin B in cell lines followed the procedure outlined above for detection of recombinant napsin A.

Napsin B was detected in tissue samples of human kidney cortex and in the human kidney cell line Hut-78: human kidney (0-40% ammonium sulfate cut); human kidney (40-70% cut); Hut-78 cells, in apparently four forms. In the 0-40% ammonium sulfate cut, a single-chain protease of 50-54 kDa with a heterogeneous amino terminus sequence derived from the protein sequence of SPGDKPIFVPLSNYR (with other termini at Asp4 and Lys5) was detected. These N-terminal sequences agreed well with the predicted activation cleavage site in pronapsin B by comparing to the activation cleavage sites in homologous procathepsin D and other aspartic protease zymogens. In the 40-70% ammonium sulfate cut, three forms were detected. A 46-50 kDa single chain form, and two two-chain forms. The 46-50 kDa band produced the same heterogeneous sequence Napsin B sequence as obtained for the larger molecular weight band in the 40% ammonium sulfate cut. The two lower molecular weight fragments of approximately 8 and 4 kDa produced the same amino-terminal sequence (VRLCLSGFQALDVPPPAGPF) corresponding to the C-terimal region of Napsin B. A prominent 40 kDa band of the transblotted preparation was sequenced, and produced the same heterogeneous amino terminal sequence as the 46-50 kDa band, indicating two species of two-chain Napsin B: an 8 kDa and 40 kDa as well as a 4 kDa and a 40 kDa species.

III. Applications of Napsin.

A variety of clinical and diagnostic uses for the enzyme can be designed based on analogy to the uses of the related aspartic proteases.

The proteins, nucleotide molecules, and methods for isolation and use thereof have a wide variety of applications, particularly in diagnostic applications. Since aspartic proteases are well known to be correlated with certain disorders, such as breast cancer and high blood pressure, and napsin is expressed in the kidney, measurement of the levels and/or types of napsin expressed in tissue, especially kidney, can be correlated with the presence and severity of disorders. The recombinant DNA and reagents derived thereform can be used to assay for napsin expression in healthy and in people inflicted with illness. Napsin sequences can be used to track the presence of napsin genes in patients for possible linkage to diseases.

A. Diagnostic Applications

The amount of napsin can be determined using standard screening techniques, ranging from isolation of napsin from the tissue, using for example immobilized anti-napsin (or anti-napsin A or anti-napsin B) or pepstatin, to detection and quantification with labelled antibodies, to determination of the amount of mRNA transcribed in the tissue, using labelled nucleotide probes.

Antibody Production

Polyclonal antibodies were produced using standard techniques for immunization of an animal with purified protein in combination with an adjuvant such as Freunds' adjuvant. Monoclonal antibodies can also be prepared using standard techniques, for example, by immunizing mice until the antibody titer is sufficiently high, isolating the spleen and doing a fusion, and then screening the hybridomas for those producing the antibodies of interest. These can be antibodies reactive with any napsin, or reactive with napsin A but not B and vice versa.

Humanized antibodies for therapeutic applications, and recombinant antibody fragments can also be generated using standard methodology. A humanized antibody is one in which only the antigen-recognition sites or complementarity-determining hypervariable regions (CDRs) are of non-human origin, and all framework regions (FR) of

variable domains are products of human genes. In one method of humanization of an animal monoclonal anti-idiotypic antibody, RPAS is combined with the CDR grafting method described by Daugherty *et al.*, Nucl. Acids Res., 19:2471-2476 (1991). Briefly, the variable region DNA of a selected animal recombinant anti-idiotypic ScFv is sequenced by the method of Clackson, T., *et al.*, Nature, 352:624-688 (1991). Using this sequence, animal CDRs are distinguished from animal framework regions (FR) based on locations of the CDRs in known sequences of animal variable genes. Kabat, H.A., *et al.*, Sequences of Proteins of Immunological Interest, 4th Ed. (U.S. Dept. Health and Human Services, Bethesda, MD, 1987). Once the animal CDRs and FR are identified, the CDRs are grafted onto human heavy chain variable region framework by the use of synthetic oligonucleotides and polymerase chain reaction (PCR) recombination. Codons for the animal heavy chain CDRs, as well as the available human heavy chain variable region framework, are built in four (each 100 bases long) oligonucleotides. Using PCR, a grafted DNA sequence of 400 bases is formed that encodes for the recombinant animal CDR/human heavy chain FR protection. The expression of recombinant CDR-grafted immunoglobulin gene is accomplished by its transfection into human 293 cells (transformed primary embryonic kidney cells, commercially available from American Type Culture Collection, Rockville, MD 20852) which secrete fully grafted antibody. *See, e.g.*, Daugherty, B.L., *et al.*, Nucl. Acids Res., 19:2471-2476, 1991. Alternatively, humanized ScFv is expressed on the surface of bacteriophage and produced in *E. coli* as in the RPAS method described below.

Pharmacia's (Pharmacia LKB Biotechnology, Sweden) "Recombinant Phage Antibody System" (RPAS) may be used for this purpose. In the RPAS, antibody variable heavy and light chain genes are separately amplified from the hybridoma mRNA and cloned into an expression vector. The heavy and light chain domains are co-expressed on the same polypeptide chain after joining with a short linker DNA

which codes for a flexible peptide. This assembly generates a single-chain Fv fragment (ScFv) which incorporates the complete antigen-binding domain of the antibody. Using the antigen-driven screening system, the ScFv with binding characteristics equivalent to those of the original monoclonal antibody is selected [*See, e.g.*, McCafferty, J., *et al.*, Nature, 348:552-554 (1990); Clackson, T., *et al.*, Nature, 352:624-688 (1991). The recombinant ScFv includes a considerably smaller number of epitopes than the intact monoclonal antibody, and thereby represents a much weaker immunogenic stimulus when injected into humans. An intravenous injection of ScFv into humans is, therefore, expected to be more efficient and immunologically tolerable in comparison with currently used whole monoclonal antibodies [Norman, D.J., *et al.*, Transplant Proc., 25, suppl. 1:89-93 (1993).

Nucleotide Probes

Nucleotide probes can be used to screen for napsin expression or the types and/or ratios of isoforms present. These can be cDNA sequences or other molecules designed based on the sequences reported herein, or which are obtained using standard techniques from libraries generated from different cell types or species. It is understood that while the sequence reported here is of human origin, the same proteases will be present in other species of animals, and will vary to some degree in both the amino acid sequence and the nucleotide sequence. Napsin is referred to herein as an aspartic protease having the naturally occuring amino acid sequence from human or other animals, or a composite sequence constructed by substitution of amino acids from one species into another, at the equivalent position, other than at the active site, discussed above. A nucleotide molecule encoding napsin can be naturally occurring, as described herein, or designed and made synthetically based on the amino acid sequence. Moreover, since at least two isoforms have been identified, it is expected that additional isoforms will be found in tissues other than kidney or liver. These isoforms are intended to encompassed within the term "napsin".

Nucleotide molecules can be used to assay for amount, type or a combination thereof, using standard diagnostic techniques. In general, probes will include a segment from a DNA encoding napsin of at least fourteen nucleotides, which should be sufficient to provide specificity under standard hybridization conditions, and even more so under stringent conditions. Reaction conditions for hybridization of an oligonucleotide probe or primer to a nucleic acid sequence vary from oligonucleotide to oligonucleotide, depending on factors such as oligonucleotide length, the number of G and C nucleotides, and the composition of the buffer utilized in the hybridization reaction. Moderately stringent hybridization conditions are generally understood by those skilled in the art as conditions approximately 25°C below the melting temperature of a perfectly base-paired double-stranded DNA. Higher specificity is generally achieved by employing incubation conditions having higher temperatures, in other words more stringent conditions. In general, the longer the sequence or higher the G and C content, the higher the temperature and/or salt concentration required. Chapter 11 of the laboratory manual of Sambrook et al., MOLECULAR CLONING: A LABORATORY MANUAL, second edition, Cold Spring Harbor Laboratory Press, New York (1990), describes hybridization conditions for oligonucleotide probes and primers in great detail, including a description of the factors involved and the level of stringency necessary to guarantee hybridization with specificity. Below 10 nucleotides, hybridized systems are not stable and will begin to denature above 20°C. Above 100,000 nucleotides, one finds that hybridization (renaturation) becomes a much slower and incomplete process, as described in greater detail in the text MOLECULAR GENETICS, Stent, G.S. and R. Calender, pp. 213-219 (1971). Ideally, the probe should be from 20 to 10,000 nucleotides. Smaller nucleotide sequences (20-100) lend themselves to production by automated organic synthetic techniques. Sequences from 100-10,000 nucleotides can be obtained from appropriate restriction endonuclease treatments. The labeling of the smaller probes with the relatively bulky

chemiluminescent moieties may in some cases interfere with the hybridization process.

Labels

Both antibodies and nucleotide molecules can be labelled with standard techniques, for example, with radiolabels, fluorescent labels, chemiluminescent labels, dyes, enzymes, and other means for detection, such as magnetic particles. For example, selective labeling of the active site with fluorescein can be performed by the method of Bock (Bock, P.E. (1988) Biochemistry *27*, 6633-6639). In brief, a blocking agent is reacted with enzyme for 1 hour at room temperature. After dialysis, the covalently modified enzyme is incubated at room temperature for one hour with 200 μM 5-(iodoacetamido)fluorescein (Molecular Probes). Free fluorescein is removed by gel filtration on a PD-10 column (Pharmacia). With this method, each molecule of fluoresceinated enzyme contains a single dye at the active site and hence all of the fluorescent molecules behave identically. Alternatively, iodogen (Pierce) can be used to radiolabel enzyme with Na[^{125}I] (Amersham) according to the manufacture's protocol. Free ^{125}I can be removed by gel filtration on a PD-10 column.

Recombinant Protein

Recombinant proteins, and fragments thereof, are useful as controls in diagnostic methods. The cDNA and gene sequences of napsin A were determined. The DNA was expressed in a recombinant system (human cell line) and the activity of the enzyme characterized. The cDNA and gene sequences of napsin B were determined. The proteins can be used as standards, or as discussed below, therapeutically as aspartic proteases and in studies of enzyme behavior. The expression of recombinant proteins from a cDNA without stop codon may offer certain advantages.

Procedures for isolation of Napsin

Antibodies and nucleotide probes are primarily useful in the detection of napsin, or its isoforms. In some cases it may also be useful

to isolate the purified protein. As described above, a procedure was devised to bind napsin A and napsin B on to a pepstatin-affinity column. Immobilized pepstatin can be used to purify either naturally occurring, or recombinant, napsin, from tissues in which it is expressed, for diagnostic applications.

B. Enzyme Applications.

The aspartic proteases may be useful in applications similar to those for which cathepsin D are used. Clinically, it may be advantageous to transfect, even transiently, the gene encoding napsin to treat disorders in which the individual is deficient in the protease, or to transfect an antisense, targeted ribozyme or ribozyme guide sequence, or triple helix to prevent or decrease enzyme expression, in individuals with disorders characterized by elevated levels of enzyme.

We claim:

1. An isolated napsin.

2. The napsin of claim 1 wherein the protein is isoform A.

3. The napsin of claim 2 having the amino acid sequence of SEQ ID No. 2.

4. The napsin of claim 2 encoded by SEQ ID No. 1.

5. The napsin of claim 1 wherein the protein is isoform B.

6. The napsin of claim 5 having the amino acid sequence of SEQ ID No. 4.

7. The napsin of claim 5 encoded by SEQ ID No. 3.

8. An isolated nucleotide molecule encoding napsin.

9. The molecule of claim 8 encoding napsin A.

10. The molecule of claim 10 as depicted by SEQ ID No. 1.

11. The molecule of claim 8 encoding napsin B.

12. The molecule of claim 11 as depicted by SEQ ID No. 3.

13. The molecule of claim 8 or a portion of at least fourteen nucleotides unique to napsin labelled with a detectable label.

14. A method for isolating napsin comprising isolating the protein bound to immobilized pepstatin in an tissue extract.

15. The method of claim 14 wherein the tissue is kidney cells.

16. A method for detecting the amount or type of napsin present in a tissue comprising reacting the tissue with a labelled nucleotide molecule probe specifically hybridizing to DNA or RNA encoding napsin, or reacting the tissue with a labelled antibody specifically immunoreactive with napsin.

17. The method of claim 16 wherein the tissue is screened for the level of expression of both napsin A and napsin B.

18. The method of claim 16 wherein the amount or type of napsin present in the tissue is compared to the amount or type of napsin present in a normal control tissue.

19. An antibody specifically immunoreactive with napsin.

20. The antibody of claim 19 wherein the antibody is immunoreactive with either napsin A or napsin B.

```
1    ATGTCTCCACCACCGCTgctgCTaccCTtGCTGCTGctGCTGcCTcTGCTGAATGTGGAGCCTGCTGGGGCCACACTGATCCGGATCCcTcTTcGTCAAG  100
     M  S  P  P  P  L  L  L  P  L  L  L  L  L  P  L  L  N  V  E  P  A  G  A  T  L  I  R  I  P  L  R  Q  V

101  TCCACCCTGGACGCAGGACCCTGAACCTACTGAGGGGATGGGGAAAACCAGCAGAGCTCCCCAAGTTGGGGGCCCCATCCCCTGGGGACAAGCCTGCCTC  200
       H  P  G  R  R  T  L  N  L  L  R  G  W  G  K  P  A  E  L  P  K  L  G  A  P  S  P  G  D  K  P  A  S

201  GGTACCTCTCTCCAAATTCCTGGATGCCCAgTATTTTGGGgaAATTGggctgGGAACGCCTCCACAAAACTTCACTGTTGCCTTTGACACTGGCTCCTCC  300
       V  P  L  S  K  F  L  D  A  Q  Y  F  G  E  I  G  L  G  T  P  P  Q  N  F  T  V  A  F  D  T  G  S  S

301  AATCTCTGGGTCCCGTCCAGGAGATGCCACTTCTTCAGTGTGCCCTGCTGGTTCCACCACCGCTTCAATCCCAATGCCTcCAGCTCCTTCAAGCCCAGTG  400
       N  L  W  V  P  S  R  R  C  H  F  F  S  V  P  C  W  F  H  H  R  F  N  P  N  A  S  S  S  F  K  P  S  G

401  GGACCAAGTTTGCCATTCAGTATGGAACTGGGCgGGTAGATGGAATCCTGAGTgAGGACAAGCTGACTATtGGTGGAATCaAGGGTGCATCCGTGATTTT  500
       T  K  F  A  I  Q  Y  G  T  G  R  V  D  G  I  L  S  E  D  K  L  T  I  G  G  I  K  G  A  S  V  I  F

501  CggGgAAgcTCTGTGGGAATCCAGcctGGTCTTCACTGTTTCCCGCCCCGATGGGATATTGGGCCTCGGTTTTCCCATTCTGTCTGTGGAAGGAgTTCGG  600
       G  E  A  L  W  E  S  S  L  V  F  T  V  S  R  P  D  G  I  L  G  L  G  F  P  I  L  S  V  E  G  V  R

601  CCCCCGCTGGATGTACTGGTGGAGCAGGGGCTATTGGaTAAGCCTGTCTTCTCCTTTTACTTCaACAGGgacCCTGAAgTGGCTGATGGAgGAgAgCTGG  700
       P  P  L  D  V  L  V  E  Q  G  L  L  D  K  P  V  F  S  F  Y  F  N  R  D  P  E  V  A  D  G  G  E  L  V

701  TCcTgggGGGcTCAgACCCGGCACACTACATCCCACCCCTCACCTTcGTGCCAGTCACAGTCCCCGCCTAcTGGCAGATCCACATGGAgCGTGTgAAgGT  800
       L  G  G  S  D  P  A  H  Y  I  P  P  L  T  F  V  P  V  T  V  P  A  Y  W  Q  I  H  M  E  R  V  K  V

801  GGGCTCACGGCTgActctcTGTGCCCAgGGCTGTGCTGCCATcCTGGAtACAgGCACACCTGTcATcGTAgGACCCActgAgGAgATCCGGGCCCTGCAT  900
       G  S  R  L  T  L  C  A  Q  G  C  A  A  I  L  D  T  G  T  P  V  I  V  G  P  T  E  E  I  R  A  L  H

901  GCAGCCATTGGGGGAATCCCCTTGCTGGCTGGGGAgTacATCATCCGGTGCTCagAAATCCCAAAGcTCCCCGCAgTcTCAcTCCTCATTGGGGGGGTcT  1000
       A  A  I  G  G  I  P  L  L  A  G  E  Y  I  I  R  C  S  E  I  P  K  L  P  A  V  S  L  L  I  G  G  V  W

1001 GGTTTAATCTCACGgCCCAgGAtTAcGTCATCCAGTTTGcTCAGGGTGAcGTCCGCcTcTGCTTGTCcGGCTTCCGGgCCTTGGACATCGCTTCGCCTCC  1100
       F  N  L  T  A  Q  D  Y  V  I  Q  F  A  Q  G  D  V  R  L  C  L  S  G  F  R  A  L  D  I  A  S  P  P

1101 AGTACCTGTGTGGATCCTCGGCGACGTTTTCTTggGGGCGTATGTGACCGTCTTCGACCGCGGGGACATGAAGAGCGGCGCaCgAgTGGGAcTGGCGCGC  1200
       V  P  V  W  I  L  G  D  V  F  L  G  A  Y  V  T  V  F  D  R  G  D  M  K  S  G  A  R  V  G  L  A  R

1201 GCTCGCCCTCgCGGAgCGGACCTGGGAAGGCGCGAGACCGCGCAGGCGCAGTACCGCGGGTGCCGCCCAGGTGATGCGCATGCGCACCGGGTAGCCGAGC  1300
       A  R  P  R  G  A  D  L  G  R  R  E  T  A  Q  A  Q  Y  R  G  C  R  P  G  D  A  H  A  H  R  V  A  E  L

1301 TagcgCTACTCAGTAAAAATCCAATATTTCCATTGAAAAAAAAAAAAAAAAA  1353
       A  L  L  S  K  N  P  I  F  P  L  K  K  K  K  K  K
```

FIGURE 1

```
                -60         -50         -40         -30         -20
H-Napsin MSPPPLLLPL LLLLPLLNVE PAGATLIRIP LRQVHPGRRT LNLLRGWGK.
M-KAP    MSP...LLLL LLCLLLGNLE PEEAKLIRVP LQRTHLGHRI LNPLNGWEQ.
H-CathD  .MQPSSLLPL ALCL....LA APASALVRIP LHKFTSIRRT MSEVGGSVED

                -10          1          10          20          30
H-Napsin ...PAELPKL GAPSPGDKPA SVP..LSKFL DAQYFGEIGL GTPPQNFTVA
M-KAP    ...LAELSR. .TSTSGGNPS FVP..LSKFM NTQYFGTIGL GTPPQNFTVV
H-CathD  LIAKGPVSKY SQAVPAVTEG PIPEVLKNYM DAQYYGEIGI GTPPQCFTVV

                    40          50          60          70
H-Napsin FDTGSSNLWV PSRRCHFFSV PCWFHHRFNP NASSSFKPSG TKFAIQYGTG
M-KAP    FDTGSSNLWV PSTRCHFFSL ACWFHHRFNP KASSSFRPNG TKFAIQYGTG
H-CathD  FDTGSSNLWV PSIHCKLLDI ACWIHHKYNS DKSSTYVKNG TSFDIHYGSG

             80          90                     100         110
H-Napsin RVDGILGEDK LTI..... ....GGIKGA SVIFGEALWE SSLVFTVSRP
M-KAP    RLSGILSQDN LTI..... ....GGIHDA FVTFGEALWE PSLIFALAHF
H-CathD  SLSGYLSQDT VSVPCQSASS ASALGGVKVE RQVFGEATKQ PGITFIAAKF

            120         130         140         150         160
H-Napsin DGILGLGFPI LSVEGVRPPL DVLVEQGLLD KPVFSFYFNR DPEVADGGEL
M-KAP    DGILGLGFPT LAVGGVQPPL DAMVEQGLLE KPVFSFYLNR DSEGSDGGEL
H-CathD  DGILGMAYPR ISVNNVLPVF DNLMQQKLVD QNIFSFYLSR DPDAQPGGEL

                170         180         190         200         210
H-Napsin VLGGSDPAHY IPPLTFVPVT VPAYWQIHME RVKVGSRLTL CAQGCAAILD
M-KAP    VLGGSDPAHY VPPLTFIPVT IPAYWQVHME SVKVGTGLSL CAQGCSAILD
H-CathD  MLGGTDSKYY KGSLSYLNVT RKAYWQVHLD QVEVASGLTL CKEGCEAIVD

                220         230         240         250         260
H-Napsin TGTPVIVGPT EEIRALHAAI GGIPLLAGEY IIRCSEIPKL PAVSLLIGGV
M-KAP    TGTSLITGPS EEIRALNKAI GGYPFLNGQY FIQCSKTPTL PPVSFHLGGV
H-CathD  TGTSLMVGPV DEVRELQKAI GAVPLIQGEY MIPCEKVSTL PAITLKLGGK

                270             280         290         300         310
H-Napsin WFNLTAQDYV IQFAQGDVRL CLSGFRALDI ASPPVPVWIL GDVFLGAYVT
M-KAP    WFNLTGQDYV IQDLQSDVGL CLLGFQALDI PKPAGPLWIL GDVFLGPYVA
H-CathD  GYKLSPEDYT LKVSQAGKTL CLSGFMGMDI PPPSGPLWIL GDVFIGRYYT

                    320   326 330         340         350
H-Napsin VFDRGDMKSG ARVGLARARP RGADLGRRET AQAQYRGCRP GDAHAHRVAE
M-KAP    VFDRGDKNVG PRVGLARAQS RSTDRAERRT TQAQFFKRRP G.........
H-CathD  VFDRDN.... NRVGFAEAAR L..........  ..........  ..........

             360         370
H-Napsin LALLSKNPIF PLKKKKKK....
M-KAP    .......... ........
H-CathD  .......... ........
```

FIGURE 2A

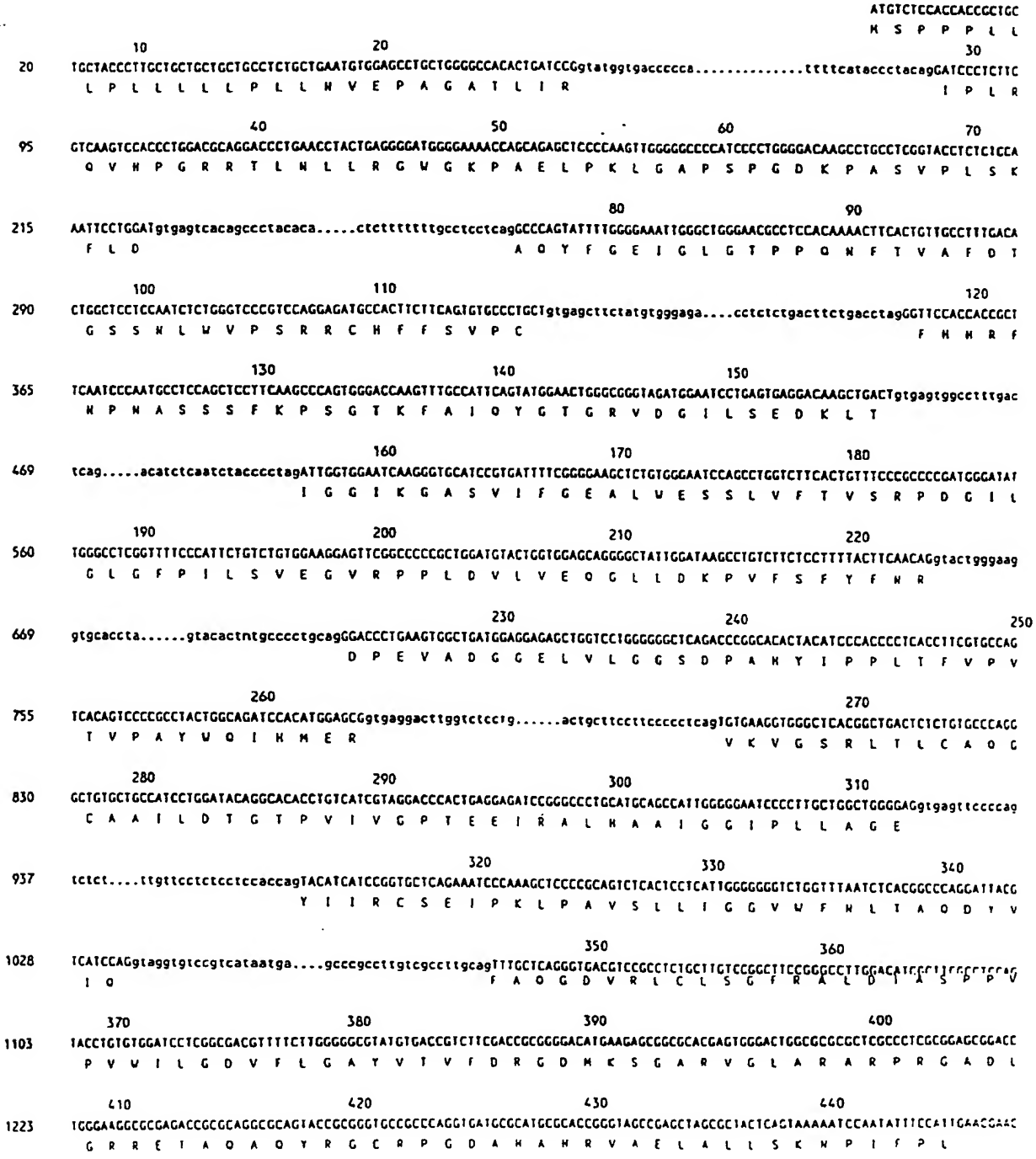

FIGURE 3A

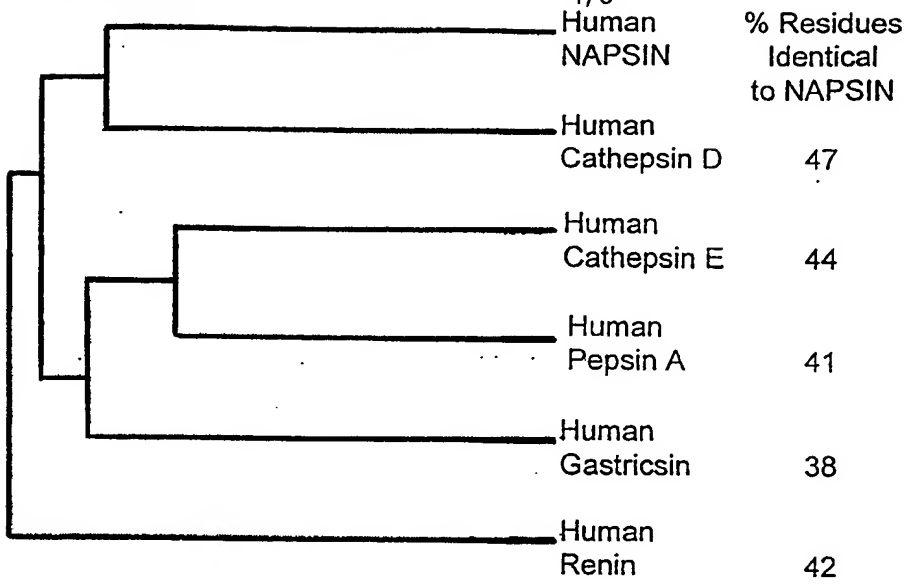

FIGURE 2B
A dendrogram presentation of sequence relatedness between hapsin and other human aspartic proteinases.

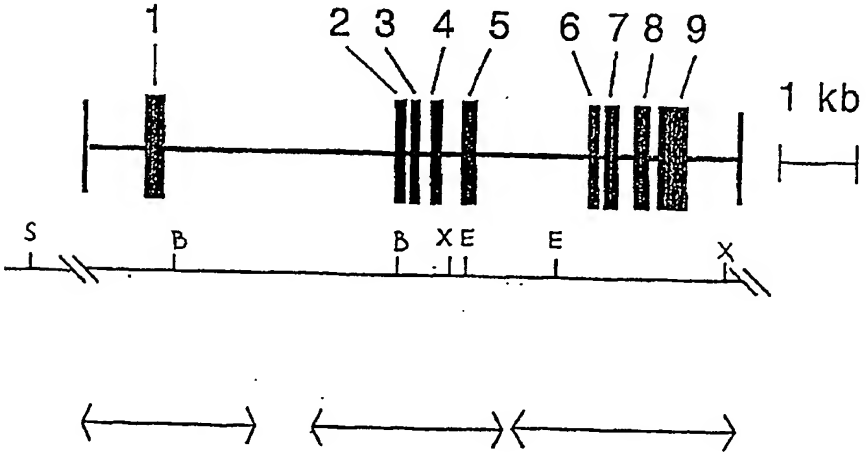

FIGURE 3B

```
1    AATGATCTGTTGTCAACAAGAAACATACTTCACCTACAAATAAAACAGTAAGAGACTGGGTCCTGAAATGCGGGCCCACTTCATATGTGAGGGCAGGTGT   100

101  CTAATCATGTCCTTTCTCCCTTCCCCCAGGCCCTTCACAGATACCTGCTGGTCTCTCCCACTTGGCCAAGGAAACANTTGTGGTTAATAAGTCTCAGAAA   200

201  AGTTATGTGAAAGTTAAAAGTAAAACTGACAGCAGCTGAAGGATGGGGGGGTGGGAGGTGGTGACGGTGGAGGAGACCCCACCACCACTGCCACCCAAGT   300

301  AGGGAGTGAGGAGCACCAGGAGCACAGGATGCTACTTCTGCCAACCCTACAAAAATACTCTGCACAAATCTTCAAAAAACATCCTTGTCCCACTGCGTCA   400

401  CCTGCGGACAGATTTCATGTCCTGGTCTCCTTCTAAACCTGGAGGTGGGGCATGAACAGGGTGGAGTCACAGGGGAAAGAAAATGAGGCCCCAGGACACC   500

501  TGGGTTCACACCCAGGTCCCCAGCGATGTCTCCACCACCGCTGCTGCAACCCCTGCTGCTGCTGCTGCCTCTGCTGAATGTGGAGCCTTCCGGGGCCACA   600
                                  M  S  P  P  P  L  L  Q  P  L  L  L  L  L  P  L  L  N  V  E  P  S  G  A  T

601  CTGATCCGCATCCCTCTTCATCGAGTCCAACCTGGACGCAGGATCCTGAACCTACTGAGGGGATGGAGAGAACCAGCAGAGCTCCCCAAGTTGGGGGCCC   700
     L  I  R  I  P  L  H  R  V  Q  P  G  R  R  I  L  N  L  L  R  G  W  R  E  P  A  E  L  P  K  L  G  A  P

701  CATCCCCTGGGGACAAGCCCATTTTCGTACCTCTCTCGAACTACAGGGATGTGCAGTATTTTGGGGAAATTGGGCTGGGAACGCCTCCACAAAACTTCAC   800
       S  P  G  D  K  P  I  F  V  P  L  S  N  Y  R  D  V  Q  Y  F  G  E  I  G  L  G  T  P  P  Q  N  F  T

801  TGTTGCCTTTGACACTGGCTCCTCCAATCTCTGGGTCCCGTCCAGGAGATGCCACTTCTTCAGTGTGCCCTGCTGGTTACACCACCGATTTGATCCCAAA   900
     V  A  F  D  T  G  S  S  N  L  W  V  P  S  R  R  C  H  F  F  S  V  P  C  W  L  H  H  R  F  D  P  K

901  GCCTCTAGCTCCTTCCAGGCCAATGGGACCAAGTTTGCCATTCAATATGGAACTGGGCGGGTAGATGGAATCCTGAGCGAGGACAAGCTGACTATTGGTG   1000
     A  S  S  S  F  Q  A  N  G  T  K  F  A  I  Q  Y  G  T  G  R  V  D  G  I  L  S  E  D  K  L  T  I  G  G

1001 GAATCAAGGGTGCATCAGTGATTTTCGGGGAGGCTCTCTGGGAGCCCAGCCTGGTCTTCGCTTTTGCCCATTTTGATGGGATATTGGGCCTCGGTTTTCC   1100
       I  K  G  A  S  V  I  F  G  E  A  L  W  E  P  S  L  V  F  A  F  A  H  F  D  G  I  L  G  L  G  F  P

1101 CATTCTGTCTGTGGAAGGAGTTCGGCCCCCGATGGATGTACTGGTGGAGCAGGGGCTATTGGATAAGCCTGTCTTCTCCTTTTACCTCAACAGGGACCCT   1200
       I  L  S  V  E  G  V  R  P  P  M  D  V  L  V  E  Q  G  L  L  D  K  P  V  F  S  F  Y  L  N  R  D  P

1201 GAAGAGCCTGATGGAGGAGAGCTGGTCCTGGGGGGCTCGGACCCGGCACACTACATCCCACCCCTCACCTTCGTGCCAGTCACGGTCCCTGCCTACTGGC   1300
     E  E  P  D  G  G  E  L  V  L  G  G  S  D  P  A  H  Y  I  P  P  L  T  F  V  P  V  T  V  P  A  Y  W  Q

1301 AGATCCACATGGAGCGTGTGAAGGTGGGCCCAGGGCTGACTCTCTGTGCCAAGGGCTGTGCTGCCATCCTGGATACGGGCACGTCCCTCATCACAGGACC   1400
       I  H  M  E  R  V  K  V  G  P  G  L  T  L  C  A  K  G  C  A  A  I  L  D  T  G  T  S  L  I  T  G  P

1401 CACTGAGGAGATCCGGGCCCTGCATGCAGCCATTGGGGGAATCCCCTTGCTGGCTGGGGAGTACATCATCCTGTGCTCGGAAATCCCAAAGCTCCCCGCA   1500
     T  E  E  I  R  A  L  H  A  A  I  G  G  I  P  L  L  A  G  E  Y  I  I  L  C  S  E  I  P  K  L  P  A

1501 GTCTCCTTCCTTCTTGGGGGGGTCTGGTTTAACCTCACGGCCCATGATTACGTCATCCAGACTACTCGAAATGGCGTCCGCCTCTGCTTGTCCGGTTTCC   1600
     V  S  F  L  L  G  G  V  W  F  N  L  T  A  H  D  Y  V  I  Q  T  T  R  N  G  V  R  L  C  L  S  G  F  Q

1601 AGGCCCTGGATGTCCCTCCGCCTGCAGGGCCCTTCTGGATCCTCGGTGACGTCTTCTTGGGGACGTATGTGGCCGTCTTCGACCGCGGGGACATGAAGAG   1700
       A  L  D  V  P  P  P  A  G  P  F  W  I  L  G  D  V  F  L  G  T  Y  V  A  V  F  D  R  G  D  M  K  S

1701 CAGCGCCCGGGTGGGCCTGGCGCGCGCTCGCACTCGCGGAGCGGACCTCGGATGGGGAGAGACTGCGCAGGCGCAGTTCCCCGGGTGACGCCCAAGTGAA   1800
     S  A  R  V  G  L  A  R  A  R  T  R  G  A  D  L  G  W  G  E  T  A  Q  A  Q  F  P  G

1801 GCGCATGCGCAGCGGGTGGTCGCGGAGGTCCTGCTACCCAGTAAAAATCCACTATTTCCATTGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA   1900

1901 AAAAAAAAAA  1910
```

FIGURE 4